DOCUMENT RESUME

ED 179 562 TM 009 829

AUTHOR Sands, William A.
TITLE Computer Assisted Reference Locator (CARL) System: An Overview.
PUB DATE 78
NOTE 22p.: Paper presented at the Annual Meeting of the Military Testing Association (20th, Oklahoma City, OK, October 30-November 3, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bibliographic Citations; Cataloging; Computer Programs; *Electronic Data Processing; *Indexing; *Information Retrieval; Information Storage; *Information Systems; Search Strategies
IDENTIFIERS *Computer Assisted Reference Locator; Computer Searches

ABSTRACT

The Computer Assisted Reference Locator (CARL) is a computer-based information retrieval system which uses coordinate indexing. Objectives established in designing the system are: (1) simplicity of reference query and retrieval; (2) ease of system maintenance; and (3) adaptability for alternative computer systems. The source documents input into the system are not limited to published materials but could include such documents or files as lecture notes, documentation for computer programs, printed advertisements, equipment brochures, print or slide collections, and record or tape collections. The retrieval system could be adapted for home or personal use as well as for professional use. Data are input by keypunching, and include four types of entries: (1) header--author, year of publication, and first letter of title; (2) bibliographic citation; (3) author; and (4) subject keywords which are based on a controlled vocabulary and may include eighteen characters in length. The computer programs used for input preparation provide editing, sorting of cards, duplicating, and printing of keyword lists. Three systems programs are used for adding to the data base, query-retrieval, and change/maintenance. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED179562

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Computer Assisted Reference Locator (CARL) System
An Overview[1]

by

William A. Sands

Acquisition and Initial Service Program
Navy Personnel Research and Development Center
San Diego, California 92152

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
William Sands

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 009 827

The 20th Annual Military Testing Association Conference

Oklahoma City, Oklahoma

30 October - 3 November 1978

[1]The opinions or assertions contained herein are those of the writer and are not to be construed as official or reflecting the views of the Navy Department.

INTRODUCTION

Background

The problem which originally prompted the present author's interest in the field of information retrieval is humorously related by Redican (1973) in an article entitled "Reprints: File Before They Defile You," published in the American Psychologist. He states that:

> Several centuries ago, a philosopher, whose name now escapes me, is said to have come to an untimely end when his overloaded bookcase toppled over and buried him. Although the condition of most psychologist's bookshelves is undoubtedly not quite so lethal, many are burdened by sizable piles of reprints, manuscripts, journals, and books that await filing (July 1973, p. 625).

The creation, utilization, and maintenance of a reference retrieval system was identified as a professional problem for the field of psychology twenty-five years ago by Daniel and Louttit (1953). The magnitude of the problem can be appreciated by the recognition that the growth rate of scientific publications is exponential, with the number of references doubling every 13-15 years (Price, 1961, 1963). This rate of information production can be overwhelming to the individual researcher who attempts to keep up with the literature in a particular field. The initial attempt at imposing order on a growing personal reference library often involves filing documents alphabetically by the last name of the senior author. Inevitably there arises a situation where the researcher knows that the library contains a reference on a particular topic but cannot recall the author's name. Location of the reference requires a sequential search of the alphabetical file of source documents. With a small reference library, this sequential search process is inconvenient. As time passes and the size of the reference library grows, the sequential search strategy becomes increasingly burdensome. Therefore, some alternative information retrieval method is sought.

Information Retrieval Systems

Lancaster (1968), adopting a broad perspective, maintains that information retrieval encompasses all the activities from the initial acquisition and indexing of source documents to the search, retrieval, and delivery of the results of a query to the user. He further points out that an information retrieval system does not change the knowledge of the user on any subject. Rather, the system simply informs the user of the presence or absence of source documents on a particular topic and the location of all pertinent documents.

A wide variety of information retrieval systems is available (Bourne, 1963; Lancaster, 1968) including a manual system using ordinary index cards, an edge-punch card system employing thin rods to sort and retrieve information, and various computer-based systems. One manual

approach which has considerable merit is the "accession number coordinated system" described by de Alarcón (1969), and, specifically, a variation known as "Uniterm" introduced by Taube (1953) and discussed in the American Psychologist by Broadhurst (1962). The coordinate indexing method advocated by Broadhurst is described as follows:

> The procedure involves setting up a separate card file to index the collection of references, these being filed irrespective of author or content but merely according to a serial number assigned to each reference as it is added--technically the accession number. The classification of the reference is then done by selecting certain key words and underlining them ('tracing'). This is most conveniently done as the reference is read, and can be done on the card or the associated reprint if you have one, or in your copy of the journal, so long as the serial number given to either of the latter is the same as that shown on the reference card. These key words (or 'Uniterms') which have been created by the tracing procedure can be as many or as few as you like, depending on your interests and the relevance of the material to them. If an important classification does not occur in the text or reference, it can be added. The serial number of the reference is then transferred ('posted') to cards which merely bear the appropriate Uniterms as headings. Proceeding in this way generates a personal psychological vocabulary of Uniterms represented by cards each having the serial numbers of references on them. Retrieval of information then becomes simple. Consideration of any one Uniterm card will give the serial numbers of all the references which deal with the subject in question. For two subjects, take the two Uniterm cards, and compare them for coincidences of number. Such coincidences of number indicate references which deal with both subjects. Making a series of such cross matches of Uniterm cards will yield information about the collection of references in as broad or as fine a detail as is required. (1962, p. 137).

Thus the term coordinate indexing is quite descriptive. Each Uniterm can be considered as one of a set of classification coordinates. For any pair of Uniterms, the accession number(s) at which the pair intersects represents a reference(s) that has been classified as belonging under both Uniterms.

THE CARL SYSTEM

Design Objectives

The Computer Assisted Reference Locator (CARL) System is a computer-based information retrieval system which generally follows the coordinate indexing approach described above. Some of the objectives considered in designing the system were: (1) simplicity of reference query and

retrieval; (2) ease of system maintenance (additions, deletions, corrections, etc.); and (3) adaptability for alternative computer systems.

The first objective, query-retrieval simplicity, is primary. If a system is too complicated, troublesome or time-consuming, a researcher probably will avoid using it. From a practical standpoint, any information retrieval system which is not used is worthless, regardless of the creativity displayed in the system design or programming.

The second objective, maintenance simplicity, is also important. If the system is so complex that changes to the existing data files or the addition of new references to the system is a monumental task, the system probably would be expensive in terms of time, costs, or both. Obviously, this expense could constrain the utility of a reference retrieval system.

The third objective, adaptability, is a desirable characteristic as it will facilitate the adoption and use of the CARL System by researchers having access to a wide variety of computer systems with different operating systems, internal and external storage conventions, and input/output characteristics. The system design and the component source programs of the CARL System have been developed with this third objective in mind. The original version of the CARL System (described in this paper) is a sequential access system, as distinguished from a direct access system. Therefore, the data files which will be described below could be stored either on magnetic tape or on disk. Most computer systems include one or both of these storage media. The source programs which incorporate the processing logic of the CARL System are written in ASCII FORTRAN.[2] The use of FORTRAN should insure a wide degree of compatibility with different computers, as most systems support FORTRAN.[3]

Input Information

The information for each new reference is encoded for keypunching[4] on four different forms. The first form is for headers, as shown in Figure 1. The header card allows space for up to four authors (last name and first and middle initials), the year of publication, and the first letter of the reference title. In addition, there is space for a five-digit reference number and a one-digit card number. Zero is the card number for the header card. Each reference has only one header card, regardless of the actual number of authors. The header card is used for sorting operations in the CARL System.

---

[2]A version of the FORTRAN language which handles the full American Standard Code for Information Interchange character set.

[3]The computer system used in the development of the CARL System is a UNIVAC 1110, a general purpose, high performance, multiprocessor system employing the EXEC 8 executive system. This computer system is located at the Naval Ocean Systems Center in San Diego, California.

[4]Actually, there are no system constraints which require card input. The information could be entered via a computer terminal.

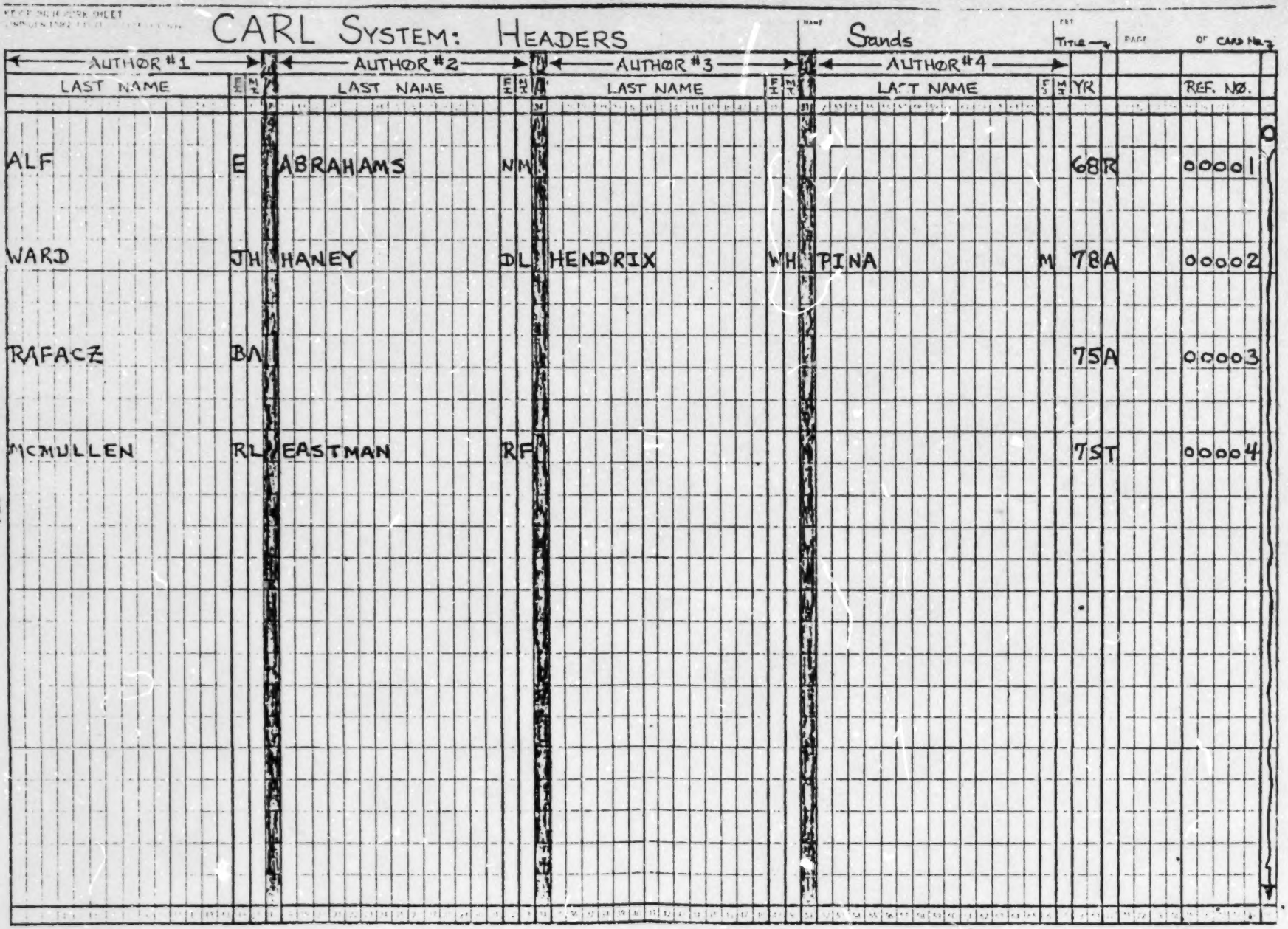

Figure 1. Header Card Keypunching Form.

The second form is used for the text of the reference citation. There are no system constraints on the arrangement of information within this text card (excluding the reference number and card number). Figure 2 illustrates the conventions adopted by the present author. The first line of text begins in column one with the senior author's last name, first and middle initials. This is followed by the article title, the journal name, year, volume number, and page numbers. Note that the additional text cards differ from the first text card in that information begins in column four. This is done to improve readability of lists of references. The second example in this figure illustrates the encoding of a technical report published by one of the military personnel research laboratories. The authors and title are formatted as above, followed by the report type and number assigned. Next, the location and the name of the performing organization are specified, followed by the publication date. The text cards for a single reference are assigned card numbers from one to six, as needed.

The third type of form is for author information. As shown in Figure 3, the format allows for four authors per card. If there are more than four authors, a second author card is used. All author cards are assigned the number seven as the card number.

The last form used to encode input data is for keywords. As shown in Figure 4, up to four keywords are allowed per card. The number of keyword cards for a single reference is unlimited, but each keyword card is assigned eight as a card number. The only system constraint on keywords is a length of eighteen characters. The present author has developed a controlled vocabulary for use in assigning keywords to references. The choice of keywords for a reference is the most crucial aspect of encoding all input data. If many keywords having only a remote relationship to the reference are assigned, the reference will be retrieved frequently when it is not useful. On the other hand, if a reference is not assigned critical keywords, it will be missed in a reference search even though the material is pertinent to the user's needs. The best balance between these two considerations will depend upon the typical user of the system. The present author leans toward overinclusion (i.e., too many keywords), to insure that all pertinent references are identified in a search.

New References

The addition of new references to an existing library involves twelve steps:

1. Alphabetize a new set of source documents by author(s) and remove any duplicates.

2. Check new source documents against existing library to identify apparent duplicates.

3. Determine unique/duplicate status of potential duplicates and eliminate identified duplicates.

KEY PUNCH WORK SHEET
FIND-GEM-10462 7-15-71-0100 LF 069-1600

CARL SYSTEM: REFERENCES

NAME: Sands | EXT | PAGE | OF CARD NO.

REFERENCE TEXT	REF. NO.
ALF, E. AND ABRAHAMS, N.M. RELATIONSHIP BETWEEN PERCENT OVERLAP AND MEASURES OF CORRELATION. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1968, 28, 779-792.	00001
WARD, J.H. JR., HANEY, D.L., HENDRIX, W.H., AND PINA, M. ASSIGNMENT PROCEDURES IN THE AIR FORCE PROCUREMENT MANAGEMENT INFORMATION SYSTEM. TECHNICAL REPORT AFHRL-TR-78-30. BROOKS AIR FORCE BASE, TEXAS: AIR FORCE HUMAN RESOURCES LABORATORY, JULY 1978.	00002
RAFACZ, B.A. A SURVEY OF THE DEVELOPMENTAL RESEARCH FOR PROJECT CONTRACT (COMPUTERIZED NAVY TECHNIQUES FOR RECRUIT ASSIGNMENT, COUNSELING, AND TESTING). PROCEEDINGS OF THE 17TH ANNUAL CONFERENCE OF THE MILITARY TESTING ASSOCIATION. FORT BENJAMIN HARRISON, INDIANA: U.S. ARMY ENLISTED EVALUATION CENTER, SEPTEMBER 1975.	00003
MCMULLEN, R.L. AND EASTMAN, R.F. THE CURRENT PREDICTIVE VALIDITY OF THE FLIGHT APTITUDE SELECTION TESTS. PROCEEDINGS OF THE 17TH ANNUAL CONFERENCE OF THE MILITARY TESTING ASSOCIATION. FORT BENJAMIN HARRISON, INDIANA: U.S. ARMY ENLISTED EVALUATION CENTER, SEPTEMBER 1975.	00004

476

9

Figure 2. Text C[illegible] Keypunching Form.

CARL SYSTEM: AUTH(S)

NAME Sands

AUTHOR #1	AUTHOR #2	AUTHOR #3	AUTHOR #4	REF. NO.
ALF E	ABRAHAMS NM			00001
WARD JH	HANEY DL	HENDRIX WH	PINA M	00002
RAFACZ BA				00003
MCMULLEN RL	EASTMAN RF			00004

Figure 3. Author Card Keypunching Form.

CARL SYSTEM: KEYWORDS

NAME: Sands

KEYWORD	KEYWORD	KEYWORD	KEYWORD	REF. NO.
RELATIONSHIP	PERCENT	OVERLAP	MEASURE	00001
CORRELATION	STATISTICS	USN	NPRDC	00001
ASSIGNMENT	PROCEDURE	PROCUREMENT	MANAGEMENT	00002
INFORMATION	SYSTEM	MIS	USAF	00002
AFHRL				00002
SURVEY	DEVELOPMENT	PROJECT	CONTRACT	00003
COMPUTER	TECHNIQUE	RECRUIT	ASSIGNMENT	00003
COUNSELING	TESTING	PROCEEDINGS	MTA	00003
USN	NPRDC			00003
PREDICTIVE	VALIDITY	FLIGHT	APTITUDE	00004
SELECTION	TEST	FAST	PROCEEDINGS	00004
MTA	USA	BESRL		00004

478

Figure 4. Keyword Card Keypunching Form.

4. Code information on new documents onto keypunch forms:

   A. Header form for Card #0
   B. Text form for Cards #1-6
   C. Author(s) form for Card #7
   D. Keyword(s) form for Card #8

5. Keypunch new information (Cards #0-8) for each new source document.

6. Verify keypunching of new information.

7. Sort new card deck by document number and card number.

8. Run edit program on sorted deck.

9. Correct any problems identified by edit program.

10. Proofread edit program output:

   A. Header
   B. Text
   C. Author(s)
   D. Keyword(s)

11. Correct any problems identified by proofreading.

12. Input corrected card deck and update system.

Data Files

The CARL System data files may be divided into those primary files accessed by the system during normal processing and those backup files kept as insurance against disaster.

There are four primary data files in the CARL System.[5]

1. RDXXXX.--The raw data file which contains all the raw data from the header, text, author, and keyword cards for each reference (i.e., cards zero through eight), arranged sequentially by reference number and card number.

2. KAXXXX.--The keyword-author file which contains all of the keywords and authors with the associated reference numbers, arranged alphabetically by keyword/author.

---

[5]The XXXX portion of each data file name is a number indicating the highest reference number currently incorporated into the system. For example, RD0400. would indicate a data file based upon references 1-400, inclusive.

3. KDXXXX.--The keyword dictionary file containing all keywords employed in the existing library, arranged in alphabetical sequence.

4. ADXXXX.--The author dictionary file containing all the authors employed in the existing library, arranged in alphabetical sequence.

All four primary data files are backed up on magnetic tape. Specifically, the raw data file (RDXXXX.) and the keyword-author file (KAXXXX.) are kept on one reel and the two dictionary files (KDXXXX. and ADXXXX.) are kept on another reel. Finally, the original punched cards are saved so that the entire CARL System could be rebuilt if all the disk files and magnetic tapes were destroyed.

Computer Programs

The computer programs in the CARL System can be divided into two categories: (1) input preparation programs, and (2) system programs. As the name implies, the purpose of the input preparation programs is to clean up the input data prior to incorporating it into the CARL System. The EDIT program reads the punched card deck (or a disk file) containing all the raw data and prints it out in a format which facilitates visual editing. In addition, the program checks for the following problems: (1) cards missing or not in sequence within a reference, (2) reference numbers not in sequence, (3) missing reference numbers, (4) duplicate reference numbers, and (5) illegal reference numbers. Appropriate diagnostic messages are printed as error conditions are encountered. Finally, after processing all the input data, the number of errors flagged and the number of missing references are reported.

The duplicate identification program (DUPL) reads the raw data file (RDXXXX.) and strips off all zero cards. These header cards are then sorted on four fields in the following order of significance: (1) author, (2) publication date, (3) first letter of title, and (4) reference number. A sorted listing of the header cards is produced. Optional outputs include: (1) identification and listing of all potential duplicate references (i.e., references with identical header cards) and (2) a comparison of potential duplicate references with a stored list of apparent duplicates which have been previously identified as unique references and a listing of the remaining potential duplicates.

The recommended keyword dictionary program (RKWD) reads a punched card deck of keywords, arranges them in alphabetical order, and prints out a dictionary of recommended keywords. This dictionary is designed to provide initial guidance for indexers as a system is getting started and includes only a few references. Later, after a substantial number of references has been indexed and incorporated into the system, a dictionary of actual keywords will be used.

There are three systems programs: (1) a system creation program (BUILD), (2) a query-retrieval program (QUERY), and (3) a system maintenance program (CHANGE). The BUILD program is used only once (assuming

the data files in the CARL System are not destroyed). As shown in Figure 5, the BUILD program reads the raw data either from the original punched card deck or a disk file containing the same information. The data are edited and if an abortive error condition is encountered a message is printed out and processing terminates.[6] If no abortive condition is found, the BUILD program creates four output files: (1) the raw data file (RDXXXX.), (2) the keyword-author file (KAXXXX.), (3) the keyword-dictionary file (KDXXXX.), and (4) the author-dictionary file (ADXXXX.). Each of these four output files can be either a disk file or a magnetic tape file, depending upon the computer system hardware available and the costs of different storage modes. At present, a good configuration appears to be having RDXXXX. and KAXXXX. created as disk files and KDXXXX. and ADXXXX. created as magnetic tape files. The first two data files will be needed almost everytime the CARL System is used but the other two files are needed only when a dictionary (either keyword or author) is required or when certain types of corrections are being made using the CHANGE program. Finally, separate printed dictionaries are produced for keywords and authors.

As shown in Figure 6, the query-retrieval program (QUERY) uses the keyword-author file (KAXXXX.) to respond to demand terminal queries from a user. The desired keyword(s) is (are) typed on a terminal by the user. The QUERY program examines the KAXXXX. data file to identify all references with the appropriate keyword(s). A message indicating the number of references located is sent to the user on the terminal. The user is given the option of having the actual reference numbers listed or the entire text of the reference citation(s) listed. If the user elects to have the entire reference citation(s) listed, the RDXXXX. data file is required. If, on the other hand, the number of references initially indicated is too many, the user can narrow the scope of the search by specifying additional keywords.

The maintenance program (CHANGE) requires access to all four data files (RDXXXX., KAXXXX., KDXXXX., and ADXXXX.) as shown in Figure 7. If the change desired involves one or more corrections to the existing system, the appropriate data files are updated. If the maintenance activity involves adding new references to the system, new raw data are input (from cards or disk), certain editing checks are performed and all four data files are updated. Finally, the operator has the option of obtaining a post-change listing of all references changed (or added).

Query-Retrieval Example

The following example is presented to illustrate the way in which a user would interact with the CARL System. After contacting the person managing the system, the user would examine a controlled dictionary of all allowable keywords. This would enable the user to formulate the retrieval request in a manner which will be meaningful to the system.

---

[6] The operational definition of an abortive error condition can be specified by the system manager.

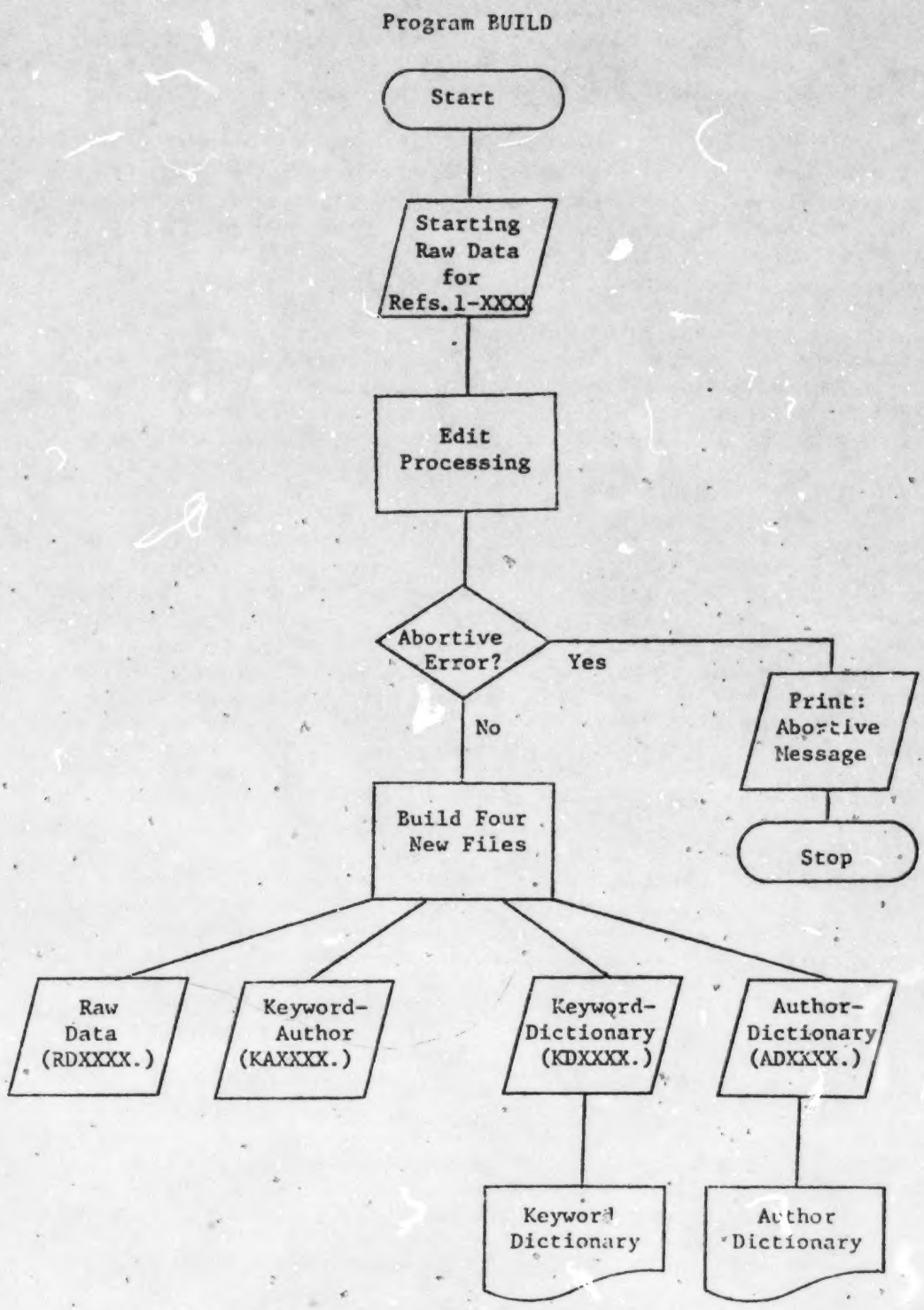

Figure 5. Program BUILD Flowchart.

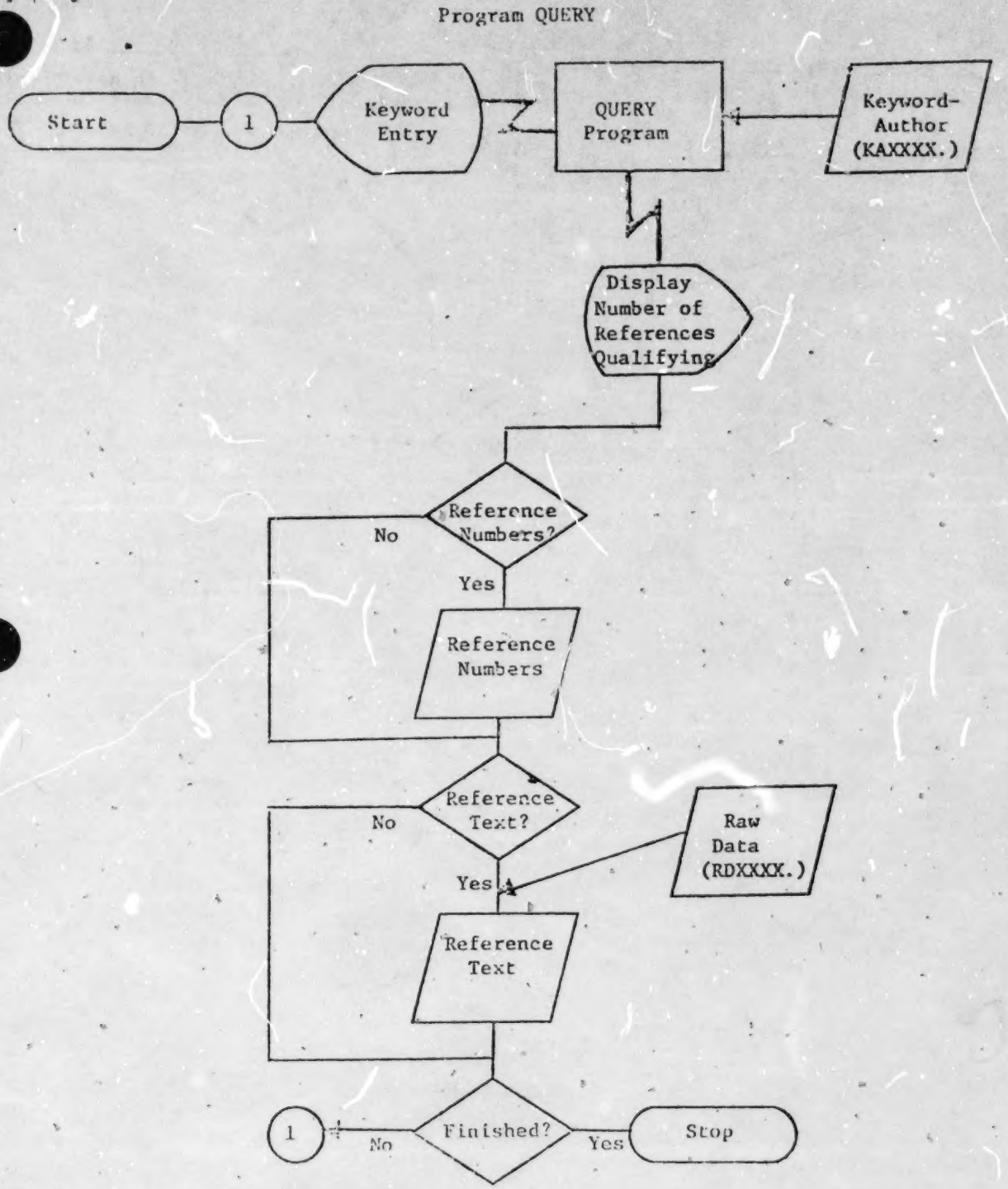

Figure 6. Program QUERY Flowchart.

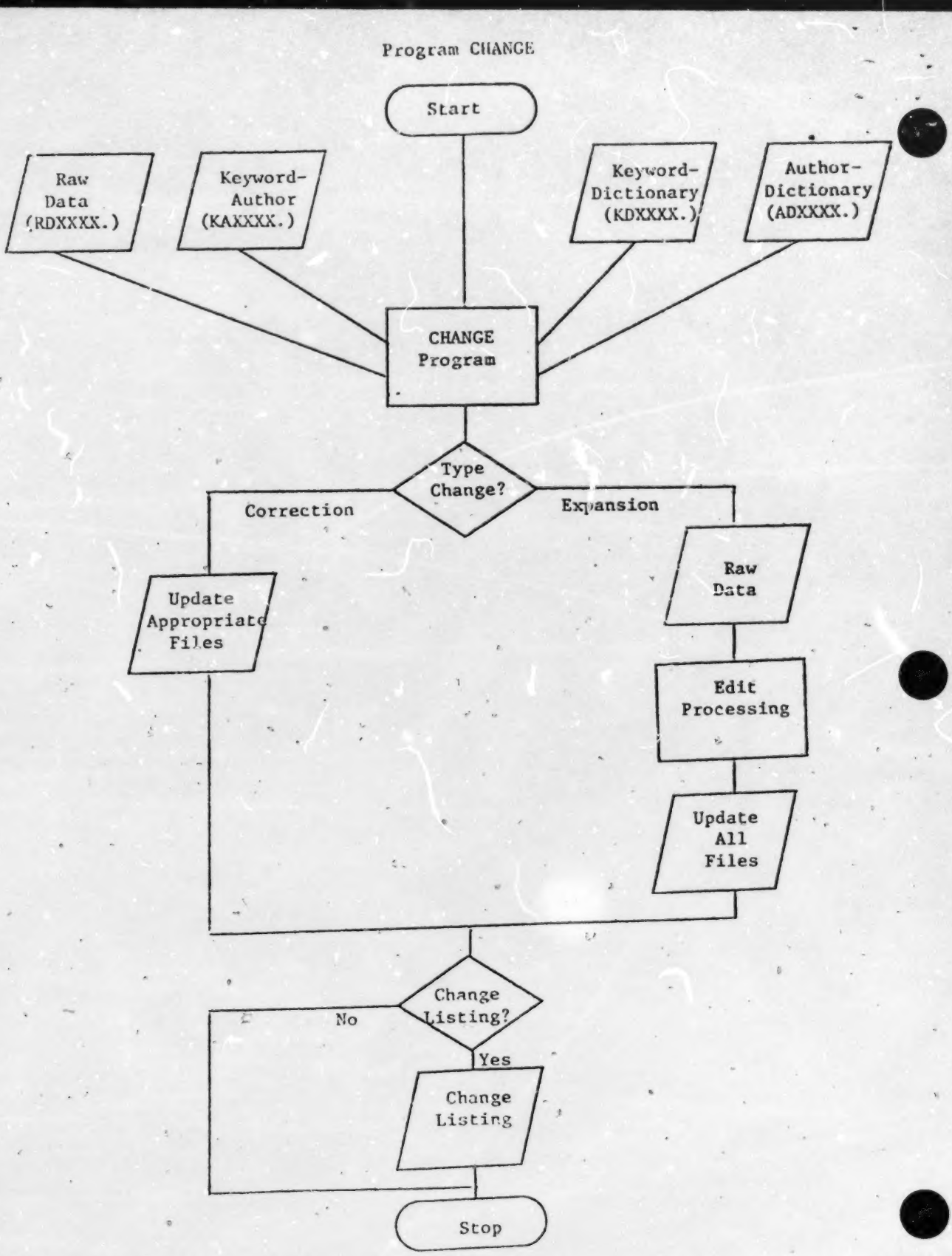

Figure 7. Program CHANGE Flowchart.

Armed with the legitimate keywords which are pertinent to the topic, the user is ready to interact with the QUERY program of the CARL System, illustrated in Figure 6. After accomplishing the log-on procedures required to establish contact with the computer system, the user would enter the keyword(s) on a demand terminal. The QUERY program will locate all references which have been indexed with that keyword, or combination of keywords, and the number of references located will be displayed on the terminal. The user is then asked if a list of reference numbers is desired and, if so, the reference numbers are printed out in ascending sequence. Next the user is provided with the option of having the entire reference citation printed out for each reference located by the search. If the user wishes the entire reference citations, a hardcopy list will be produced in alphabetical sequence by author. Finally, the user is presented with the option of continuing the interaction with the CARL System or terminating the session and logging-off the computer system.

CONCLUSIONS

Coordinate Indexing

The coordinate indexing approach to information retrieval allows considerable flexibility. The source documents contained in the personal library system are not limited to published material available to the public as is the case with commercial reference retrieval systems. Lecture notes for teaching, documentation for computer programs, printed advertisements, equipment brochures, and notes used to present briefings are examples of the diversity which can be incorporated into a personal reference system. In addition, the coordinate indexing approach has considerable generality. For example, in the work setting, coordinate indexing could be employed in a management information system (MIS) designed to provide military laboratory managers with current information on all on-going research and development projects. The source document for this MIS could be DD-1498 Forms. In a home setting, coordinate indexing could be used to create, query-retrieve, and maintain files on a minicomputer. These home files might contain a photography collection of prints or slides or a record collection of albums or tapes. Unquestionably there are many examples, both in the office and the home, where an information retrieval system employing coordinate indexing could be quite useful.

Future Work

Eventually a disk-oriented version of the CARL System will be created. This second version of the CARL System will use direct access methods as opposed to the sequential access methods used by the original CARL System described herein. The direct access version will require considerably more programming time to develop than the sequential access version. However, once developed and debugged, the direct access version should significantly speed up the information retrieval process while simultaneously providing a substantial reduction in computer costs. Further, the advantage of a direct access system over a sequential access system will increase as the number of references in the system increases.

The possibility of a CARL System Network will be given consideration. This network would allow individual researchers to have access to the personal reference libraries of other researchers, in a mutually agreed upon fashion, thereby increasing the number of references examined for any query.

REFERENCES

Bourne, C. P. Methods of Information Handling. New York: John Wiley & Sons, Inc., 1963.

Broadhurst, P. L. Coordinate Indexing: A Bibliographic Aid. American Psychologist, 1962, 17, 137-142.

Daniel, R. S., & Louttit, C. M. Professional Problems in Psychology. New York: Prentice Hall, Inc., 1953.

de Alarcón, R. A Personal Medical Reference Index. The Lancet, 1969, 1, 301-305.

Lancaster, F. W. Information Retrieval Systems: Characertistics, Testing, and Evaluation. New York: John Wiley & Sons, Inc., 1968.

Price, D. J. S. Science Since Babylon. New Haven: Yale University Press, 1961.

Price, D. J. S. Little Science, Big Science. New York: Columbia University Press, 1963.

Redican, W. K. Reprints: File Before They Defile You. American Psychologist, 1973, 28, 625-627.

Taube, M. Studies in Coordinate Indexing, Vol. 1. Washington, D.C.: Documentation Inc., 1953.